The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
World Wide Web and Population Health Intelligence:
Technical Report WS-16-15

Discovering Relevant Hashtags for Health Concepts: A Case Study of Twitter

Quanzhi Li, Sameena Shah, Rui Fang, Armineh Nourbakhsh, Xiaomo Liu
Research and Development
Thomson Reuters
3 Times Square, NYC, NY 10036
{quanzhi.li, sameena.shah, rui.fang, armineh.nourbakhsh, xiaomo.liu}@thomsonreuters.com

Abstract

Hashtags are useful in many applications, such as tweet classification, clustering, searching, indexing and social network analysis. This study seeks to recommend relevant Twitter hashtags for health-related keywords based on distributed language representations, generated by the state-of-the-art Deep Learning technology. The word embeddings are built from billions of tweet words without supervision. To the best of our knowledge, this is the first study of applying distributed language representations to recommending hashtags for keywords. The experiment showed that this approach outperformed the baseline approach that is based on keyword and hashtag co-occurrence in tweets.

Introduction

In Twitter, a hashtag is a string of characters preceded by the # character, and is used to build a topic community or as a descriptive label (Tsur and Rappoport 2012). Usually a hashtag consists of short, often abbreviated terms, and it is hard to understand its meaning wa ithout context or definition. On the other hand, given a keyword, such as "cancer", one wants to know what existing hashtags are closely related to this keyword. Manually searching social media data to find hashtags relevant to a keyword is tedious and the result is not exhaustive. A process of automatically discovering relevant hashtags for keywords is necessary. This paper presents such a method. This study focuses on the health domain, but the proposed approach is generic enough that it can be applied in other domains.

Identifying relevant hashtags for health keywords can benefit the following health information related applications: 1. Public health tracking systems based on social media can use relevant hashtags to increase its surveillance coverage. For example, MappyHealth (http://nowtrending.hhs.gov), HealthTweets.org (Dredze et al. 2014) and Crowdbreaks (http://crowdbreaks.com) just track keywords and use tweets containing those keywords to do trending and other types of analysis. By tracking only keywords, these systems will miss many relevant tweets because lots of tweets contain related hashtags but not the keywords. 2. In a health social data search platform, users can search on hashtags relevant to a keyword, in addition to the keyword itself. 3. Hashtags can be used in automatic query expansion to increase the recall of a query, and it can also be used for query suggestion. 4. As topic surrogates, they can be used in tweet clustering and classification, and in social network analysis.

To use hashtags in the aforementioned applications, one challenge is how to automatically discover the hashtags highly relevant to a given keyword or concept. Most previous studies on finding relevant hashtags focus on recommending hashtags for a tweet, instead of a keyword (Li and Wu 2009, Zangerle et al. 2011, Godin et al. 2013, She and Chen 2014). This study focuses on recommending hashtags for keywords or concepts, not tweets.

Distributed representations of words are also called word embeddings. A word embedding is a low-dimensional, dense and real-valued vector for a word (Mikolov et al. 2013). They are usually generated from a large text corpus and the embedding of a word captures both its syntactic and semantic aspects. They help learning algorithms to achieve better performance in natural language processing (NLP) tasks by grouping similar words together, and have been used in many NLP applications. Traditional bag-of-words and bag-of-n-grams hardly capture the semantics of words, or the distances between words. This means that words "pretty," "beautiful" and "train" are equally distant in spite of the fact that semantically, "pretty" should be closer to "beautiful" than "train." Based on word embeddings, "pretty" and "beautiful" will be very close to each other. In this study, the word embedding representation model is computed using a neural network, and generated from a large corpus - billions of words from tweets - without any supervision. The learned vectors

explicitly encode many linguistic regularities and patterns, and many of these patterns can be represented as linear translations. For example, the result of a vector calculation v("Beijing") – v("China") + v("Japan") is closer to v("Tokyo") than any other word vector (Mikolov et al. 2013).

One advantage of using this approach is that it is an unsupervised process, and rebuilding the model to handle new hashtags is just a matter of ingesting new tweets to the building process periodically. It doesn't require any labeled data.

The major contributions of this study are:

1. To the best of our knowledge, this is the first study that exploits distributed representations of words to recommend hashtags for a concept.
2. The proposed approach has practical applications in health social media platforms, such as HealthTweets.org, nowtrending.hhs.gov and crowdbreaks.com. By utilizing this method, these systems can be enhenced by tracking related hashtags, in addition to health keywords.
3. We release the word embedding vectors learned from 200 million tweets and billions of words. Users can query this model to find relevant hashtags, as well as relevant keywords, for a given term. They can also find relevant keywords for a given hashtag from this model.

In the following sections, we present related studies, the research overview, evaluation data set and the experiment result.

Related Studies

Several health information platforms based on social media data have been implemented (Dredze et al. 2014, Wang et al. 2014, Paul 2015). One of their main functions is the trending analysis of certain health topics by tracking health-related keywords. HealthTweets.org is a research platform for sharing the latest developments in mining health trends from Twitter and other social media sites (Dredze et al. 2014). MappyHealth.com (http://nowtrending.hhs.gov) fetches real time data from Twitter associated with their predefined health terms, and then analyzes those tweets and their condition sets for trending analysis. Crowdbreaks (http://crowdbreaks.com) is a crowdsourced disease surveillance system that collects tweets containing disease-related keywords. All these systems are based on keywords, and do not include hashtags in their tracking terms or search indexes. And they will miss many relevant tweets because lots of tweets contain related hashtags but not the keywords.

Previous studies of recommending hashtags mainly focus on identifying relevant hashtags for tweets, not for a keyword. They exploit the similarity between tweets. Li and Wu (2009) use WordNet similarity information to recommend hashtag for a tweet. Mazzia and Juett (2009) apply Bayes' rule to estimate the maximum a posteriori probability of each hashtag given the words of the tweet. Zangerle et al. (2011) recommend hashtags based on the well-known tf.idf representation of the tweet. She and Chen (2014) treat hashtags as labels of topics, and develop a supervised topic model to discover relationship among words, hashtags and topics of tweets. They also incorporate user following relationship into their model. Latent Dirichlet Allocation is used to model the underlying topic assignment of language classified tweets in (Godin et al. 2013), using of a topic distribution to recommend general hashtags. None of these studies focuses on recommending hashtag for a keyword.

A word embedding is a dense, low-dimensional and real-valued vector for a word. And it has been researched in previous studies (Socher et al. 2014, Mikolov et al. 2013). One implementation is the word2vec from Mikolov et al. (Mikolov et al. 2013). This model has two training options, continuous bag of words (CBOW) and the Skip-gram model. Both models have been used by several previous studies, mainly in sentiment analysis applications (Mass et al. 2014, Matt 2015, Tang 2014).

Study Overview and Data Set

In this section, we first introduce the distributed representations of words, which is learned by a neural network, then the data set used to build the vector model for this study, and finally our evaluation approach.

Distributed Representations of Words

Distributed representations of words have been used in other NLP related applications, but they have not been explored in discovering hashtags for keywords. A distributed language representation X consists of an embedding for every vocabulary word in space S with dimension D, where D is the dimension of the latent representation space. The embeddings are learned to optimize an objective function defined on the original text, such as likelihood for word occurrences.

One implementation is the word2vec from Mikolov et al. (2013). This model has two training options, continuous bag of words (CBOW) and the Skip-gram model. The Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. Based on previous studies and the experiments in this study, the Skip-gram model produces better results, and here we briefly introduce it.

The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document. Given a sequence of training words $W1, W2, W3, \ldots, W_N$, the Skip-gram model aims to maximize the average log probability

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{-m\le i\le m,\ i\neq 0} log\ p(Wn+i\ |\ Wn)$$

where *m* is the size of the training context. A larger *m* will result in more training data and can lead to a higher accuracy, at the expense of the training time.

Generating word embeddings from text corpus is an unsupervised process. To get high quality embedding vectors, a large amount of training data is necessary. After training, each word, including all hashtags in the case of tweet text, is represented by a low-dimensional, dense and real-valued vector. Usually the dimension size ranges from tens to hundreds.

Tweet Data Set for Building the Vector Model

Tweets used in this study date from October 2014 to September 2015. They were acquired through Twitter's public 1% streaming API and Twitter's Decahose data (10% of Twitter streaming data) granted to us by Twitter for research purpose.

Table 1 shows the basic statistics of the data set used in this study. Only English tweets are used, and about 200 million tweets are used for building the vector model. Totally, 2.9 billion words are processed. With a term frequency threshold of 5 (tokens less than 5 occurrences in the data set are discarded), the total number of unique tokens (hashtags and words) in this model is 1.9 million. The word embedding dimension is set to 300.

Each tweet is preprocessed to get a clean version, which is then processed by the model building process. The preprocess steps are as follows:

- All urls are removed
- All mentions are removed
- Dates and years are converted to two symbols representing date and year
- All ratios, such as 3/7, are replaced by a special symbol;
- Integers and decimals are normalized to two special symbols;
- All special characters, except hashtags symbol #, are removed.

These preprocessing steps are necessary, since most of tokens removed or normalized are not useful, such as various numbers and URLs, and keeping them will increase the vector space size and computing cost. Stop words are not removed, since they provide important context in which other words are used.

Number of Tweets	198 million
Number of words in training data	2.9 billion
Number of unique tokens in the trained model	1.9 million

Table 1. Basic Data Set Statistics

Evaluation Method

Since we didn't find any previous study on recommending hashtags for keywords, we evaluated our approach by comparing it to the baseline system described below.

The Baseline

We define the baseline using the term co-occurrence method, which is a very popular approach in identifying the relationship between two entities. In this study, the relationship is defined as the relevance between a keyword and a hashtag. If the keyword and a hashtag appear in the same tweet, then they co-occur in this tweet. The hashtags are ranked according to their frequencies of concurrence with the keyword. For comparison with our approach, for each tested keyword, the top 10 hashtags are returned as the relevant ones. For each keyword, the 200 million tweets are processed to find its relevant hashtags.

Our Approach

For each tested keyword, the top 10 hashtags were generated as follows: the keyword's 300-dimensional word embeddings were obtained by querying the trained model; the cosine similarity score was calculated between this keyword's embedding vector and the embedding vector of each hashtag in this model; the 10 hashtags with the highest scores were selected. Cosine similarity is a popular measure for computing the similarity between two vectors.

Table 2. Top 10 hashtags for the term "vaccine"

Baseline		Distributed word representations	
Hashtag	**Frequency of co-occurrence**	**Hashtag**	**Cosine similarity**
#ebola	1392	#vaccine	0.763
#vaccines	575	#vaccines	0.613
#vaccine	539	#zmapp	0.499
#cdcwhistleblower	524	#influenza	0.477
#vaccineswork	330	#getvaccinated	0.471
#health	278	#rubella	0.460
#flu	277	#ebolaoutbreak	0.459
#news	248	#iamtheherd	0.459
#hearthiswell	205	#flu	0.458
#gopdebate	118	#ebolacure	0.458

Comparing the Two Approaches

65 popular health-related keywords, such as flu and cancer, were selected for the evaluation. To compare the two approaches, we took the top 5 and top 10 hashtags for each tested keyword, and compared the two methods at these two levels. Each hashtag was evaluated by two domain experts, by assigning a score from 1 to 5, with 1 meaning not-relevant and 5 meaning definitely relevant. The final

score for a hashtag is the average of scores from the two experts.

For each hashtag, we provided 15 tweets containing the hashtag to help the annotators to understand its meaning. The annotators could also check a popular hashtag definition web site, https://tagdef.com, to find its definition, and use Twitter's website to search related tweets to better understand a hashtag.

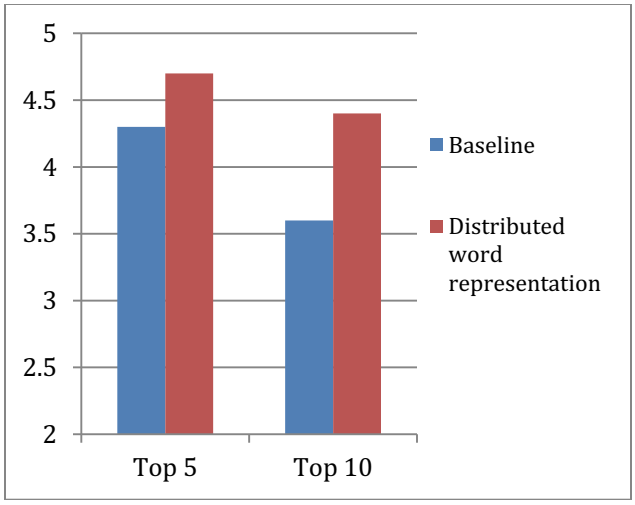

Figure 1. Performance comparison between the baseline and the distributed word representation approach

Experiment Result

We use the term "vaccine" as an example to see the top 10 hashtags returned by the two approaches. Table 2 shows the top 10 hashtags for term "vaccine". In Table 2, "Frequency of co-occurrence" is the frequency that a hashtag co-occurs with term "vaccine" in a tweet. "Cosine similarity" is the cosine similarity score between the word embedding vector of term "vaccine" and a hashtag's embedding vector. A score of 1 means the two vectors are identical and 0 means they have no relation. Tables 2 does show some difference between these two approaches. For example, hashtag #zmapp, which is a vaccine drug, is at top 3 using the proposed approach, but not in top 10 in the baseline list.

Figure 1 show the comparison between the baseline approach and our approach. It shows our approach outperformed the baseline on both the top 5 and top 10 levels. The results are statistically significant at p-value of 0.01 using paired t-test.

Conclusion

In this study, we proposed to use the distributed representations of words, which are generated by training on billions of tweet words, to discover relevant hashtags for health keywords. The experiment shows that this approach outperformed the traditional term co-occurrence based approach. The proposed method is an unsupervised learning process and can be used in any content domain. To the best of our knowledge, this is the first study exploiting distributed word representations to recommend hashtags for keywords

References

Dredze, M.; Cheng, R.; Paul, M. and Broniatowski, D., 2014. HealthTweets.org: A Platform for Public Health Surveillance using Twitter, *AAAI Workshop on the World Wide Web and Public Health Intelligence*

Godin, F.; Slavkovikj, V.; Neve, W.; Schrauwen, B. and Walle, R. 2013. Using topic models for Twitter hashtag recommendation. In *Proceeding of WWW '13 Companion*, Pages 593-596

Li, T.; Wu, Y. and Zhang, Y., 2011. Twitter hash tag prediction algorithm, *In proceeding of ICOMP'11*

Maas, A.; Daly, R.; Pham, P.; Huang, D.; Ng, A. and Potts, C., 2012. Learning word vectors for sentiment analysis, *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*

Matt, T., Document Classification by Inversion of Distributed Language Representations, 2015. *53th ACL conference*, page 45-49, July 26-31, Beijing,

Mazzia, A. and Juett, J., 2009. Suggesting Hashtags on Twitter, *technical report, Computer Science and Engineering, University of Michigan*, 2009.

Mikolov, T.; Chen, K.; Corrado, G. and Dean J., 2013. Efficient Estimation of Word Representations in Vector Space. *In Proceedings of Workshop at ICLR.*

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. and Dean J., 2013. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS

Otsuka, E.; Wallace, S. and Chiu, D. 2014. Design and Evaluation of a Twitter Hashtag Recommendation System, *IDEAS'14*, July 07 - 09 2014, Porto, Portugal

Paula, M.; Dredzea, M.; Broniatowskib, D. and Generousc, N., 2015. Worldwide Influenza Surveillance through Twitter, *AAAI Workshop on the World Wide Web and Public Health Intelligence*

She, J. and Chen, L, 2014. TOMOHA: TOpic MOdel-based HAshtag Recommendation on Twitter. *WWW'14 Companion,* April 7–11, 2014, Seoul, Korea.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.; Ng, A. and Potts, C., 2014,. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, *EMNLP 2014*

Tang, D.; Wei, F.; Yang, Y.; Zhou, M.; Liu, T. and Qin, B. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification, *52th ACL*, Baltimore, Maryland.

Tsur, O. and Rappoport, A. 2012. What's in a hashtag: content based prediction of the spread of ideas in microblogging communities. *WSDM '12*, New York, NY

Wang, S.; Paul, M. and Dredze, M., 2014. Exploring Health Topics in Chinese Social Media: An Analysis of SinaWeibo, *AAAI Workshop on the World Wide Web and Public Health Intelligence*

Zangerle, E.; Gassler, W. and Specht, G. 2011. Recommending#-tags in twitter, *in Proceedings of the Workshop on Semantic Adaptive Social Web*